

Sampling from Diffusion Networks

Motahareh Eslami Mehdiabadi*, Hamid R. Rabiee[†] and Mostafa Salehi*

Department of Computer Engineering, Sharif University of Technology

* Email: {eslami, mostafa_salehi}@ce.sharif.edu

[†] Email: rabiee@sharif.edu

Abstract—The diffusion phenomenon has a remarkable impact on Online Social Networks (OSNs). Gathering diffusion data over these large networks encounters many challenges which can be alleviated by adopting a suitable sampling approach. The contributions of this paper is twofold. First we study the sampling approaches over diffusion networks, and for the first time, classify these approaches into two categories; (1) Structure-based Sampling (SBS), and (2) Diffusion-based Sampling (DBS). The dependency of the former approach to topological features of the network, and unavailability of real diffusion paths in the latter, converts the problem of choosing an appropriate sampling approach to a trade-off. Second, we formally define the diffusion network sampling problem and propose a number of new diffusion-based characteristics to evaluate introduced sampling approaches. Our experiments on large scale synthetic and real datasets show that although DBS performs much better than SBS in higher sampling rates (16% ~ 29% on average), their performances differ about 7% in lower sampling rates. Therefore, in real large scale systems with low sampling rate requirements, SBS would be a better choice according to its lower time complexity in gathering data compared to DBS. Moreover, we show that the introduced sampling approaches (SBS and DBS) play a more important role than the graph exploration techniques such as Breadth-First Search (BFS) and Random Walk (RW) in the analysis of diffusion processes.

I. INTRODUCTION

Information diffusion as a new area of multi-disciplinary research has a remarkable effect on social networks [1]. In recent years, large Online Social Networks (OSN) such as Facebook, Twitter and YouTube have been the source of information propagation in different formats such as posts, tweets, and videos. The growth in the size of these networks results in large information networks. For example, in March 2011, Twitter users were sending 50 million tweets per day [2]. Therefore, collecting the diffusion data over large scale OSNs is often infeasible in many applications. This challenge necessitates the need for manipulating the diffusion data in an efficient way to analyze the diffusion process behavior.

Sampling strategy can be considered as a solution to solve this problem by decreasing the expense of processing on large real networks. In recent years, a considerable amount of research has been done on analyzing the topological characteristics of large OSNs based on the sampled data of different networks such as Facebook [3], Twitter [4], YouTube [5], and other large networks [6], [7]. However, considering the sampling approaches to study diffusion behaviors of social networks, apart from their topologies, is a remarkable issue that should be addressed.

To the best of our knowledge, there is no comprehensive study about sampling approaches on diffusion process. Looking into several large diffusion network studies [8], [9], [10], [11], we classify the data gathering approaches of diffusion process into two categories; (1) Structure-based Sampling (SBS), and (2) Diffusion-based Sampling (DBS). The former approach is based on the topology of the network and the latter considers propagation paths and diffusion process in sampling methodology. The SBS approach will produce some redundant data that results in decreasing the accuracy of diffusion process measurements. On the other hand, The DBS approach can reduce the redundant data and consequently increase the sampling efficiency by following the diffusion paths. However, obtaining the real diffusion paths is not practical in many applications [1], [12]. These challenges converts the problem of choosing an appropriate sampling approach for diffusion process analysis to a trade-off between many parameters such as the amount of the sampled data and the availability of diffusion paths.

In this paper, we evaluate the performance of the proposed sampling approaches to show how different sampling approaches can impact the measurement of diffusion process. To this end, our methodology comprises two steps. First, we formally define the diffusion network sampling problem. Second, we propose a number of new evaluation characteristics for the diffusion process in order to analyze their behaviors based on different sampling approaches. The proposed characteristics alleviate the dependency to topological features of the network and increase the correlation with the diffusion process. Moreover, we categorize these characteristics into three classes; (1) node-based, (2) link-based, and (3) cascade-based.

We analyze the introduced approaches by extensive experiments over large synthetic and real datasets. Two well-known sampling techniques of Breadth-First Search (BFS) and Random Walk (RW) are used in both DBS and SBS approaches. Our experiments show that the accuracy of measuring node-based and link-based characteristics in DBS grows more than SBS by increasing the sampling rate. This phenomenon will result in a considerable performance difference between these approaches in higher sampling rates (up to 65% difference). Nevertheless, cascade-based characteristics can decrease this performance difference compared to the node-based and link-based characteristics. This is the result of inherent difference between these characteristics as the formers are individual-based characteristics while the latter is related to the cascades

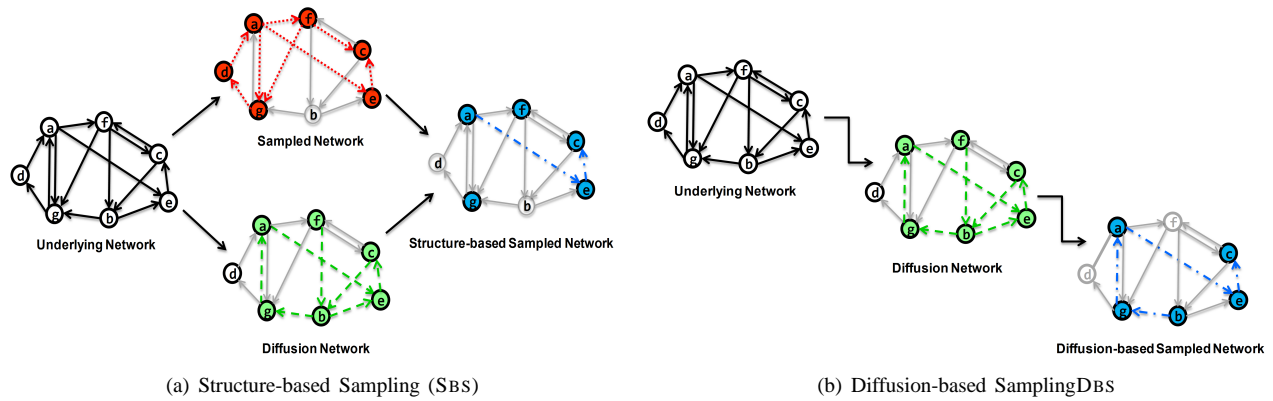


Fig. 1. Sampling Approaches. The dashed, dotted and dash-dot lines illustrate the links of diffusion, sampled and the final generated networks, respectively.

as a group-based characteristic.

Our evaluation on performance between the proposed sampling approaches shows that SBS is similar to DBS in low sampling rates (the difference is about 7% in average). The results demonstrate that SBS can be used in real systems in which only low sampling rates are feasible. Furthermore, we investigate the performance of RW and BFS in measuring diffusion characteristics. Our experiments reveal that these sampling techniques perform similar to each other (with 3% difference in average). This clarifies that the proposed sampling approaches (DBS and SBS) have more important effect than the graph exploring techniques such as RW and BFS.

In summary, our main contributions can be summarized in the following:

- Classifying and proposing sampling approaches of SBS and DBS for diffusion process analysis
- Defining the diffusion network sampling formally
- Proposing and categorizing some new diffusion-based characteristics to study the behavior of sampling approaches
- Evaluating sampling approaches (SBS and DBS) and sampling techniques (BFS and RW) in different sampling rates to evaluate their performances

The rest of the paper is organized as follows. Section II presents a classification of data collection approaches in the field of information diffusion networks. The problem definition is proposed in Section III. Section IV provides the performance evaluations, and the concluding remarks are provided in Section V.

II. DIFFUSION NETWORK DATA COLLECTION

Diffusion networks have attracted considerable attention in recent years [1], [8], [11], [12]. In spite of this great attention, there is no comprehensive survey on how to collect data from a diffusion network. The most close work to ours is [9], which studies the impact of some sampling techniques (such as Random sampling) on the information diffusion process. This work does not consider sampling approaches and their

effect on diffusion process analysis which we address in this paper. In the following, we propose a new classification of diffusion data sampling approaches.

Structure-based Sampling: The most common approach for sampling the diffusion process is to sample the underlying network and then extracting the diffusion links from the collected data (refer Figure 1(a)). Since this approach is based on the structure of the underlying network, and not the diffusion process, we call it the structure-based sampling approach (SBS). Sampling the Twitter network to study on the resulting diffusion network [9], [10] and inferring diffusion topics from the DBLP database [14] are some examples which utilize SBS to analyze the diffusion process. Using this approach will result in extraction of some redundant data such as nodes and links which do not participate in the diffusion process. Therefore, these data should be removed from the collected data to obtain the sampled diffusion network. This data reduction leads to a smaller sampled graph which may decrease the accuracy of analysis.

Diffusion-based Sampling: To study on diffusion networks, one may track the diffusion paths instead of the network paths. This idea leads to another sampling approach that explicitly considers the diffusion characteristics. We call this the diffusion-based sampling (DBS) (refer to Figure 1(b)). Recently, this approach is used in [1], [11] to collect the diffusion data.

Since diffusion network is a sub-graph of the underlying network, using DBS will increase the accuracy of the diffusion process analysis. Moreover, this approach reduces the cost of data collection by sampling only the diffusion data (i.e. the redundant data is not collected). For example, comparing SBS and DBS in Figure 1, it is observable that with the same sampling rate of 0.5 with respect to the edges, the resultant graph in DBS approach contains more links which participate in the diffusion process than the one in SBS. Nevertheless, DBS can not be used in for applications in which direct access to the links of the diffusion network is not feasible.

Actually, the latent nature of the diffusion network structure does not allow us to explore it (as simple as the underlying net-

work) [1], [12]. Therefore, choosing an appropriate sampling approach will be a trade-off between many conditions such as the amount of the sampled data and the availability of diffusion paths. Recently, we have proposed a novel sampling technique by utilizing the diffusion process characteristics [13]. In particular, it uses the infection times as local information without any knowledge about the latent structure of diffusion network.

Sampling Techniques: Since the structure of the original network is unknown initially, we use the graph exploration techniques of Breadth-First Search (BFS) and Random walk (RW) in both SBS and DBS approaches. BFS is a basic graph-based sampling technique that has been used extensively for sampling the networks in various domains [4], [5], [15]. At each iteration of BFS, the earliest explored node is selected next, and eventually, all nodes within some distance from the starting node is discovered. RW [16] is also one of the most widely used exploration sampling techniques in different kind of network contexts such as uniformly sampling Web pages from the Internet [17], degree distributions of the Facebook social graph [3] and in general large graphs [6]. A classic RW samples a graph by moving from a node u , to a neighboring node v , through an outgoing link (u, v) , chosen uniformly at random from the neighbors of node u .

III. DIFFUSION NETWORK SAMPLING

A. Preliminaries

Let $G = (V, E)$ with $n = |V|$, and $m = |E|$ be the graph representing a social network, where V is the set of nodes, and E is the set of unweighted links between pairs of nodes. Network G is called the underlying network since the information diffusion process will occur over G . Spreading some diffusible chunks of information over the underlying network creates a path which is called a cascade. A cascade can transmit some pieces of information such as epidemic diseases; Therefore, we may refer to these diffusible chunks as “infection” [12]. Each cascade c has n_c edges that is shown by an Infection Vector (IV) in which the order of edges illustrates the order of cascade passage over them:

$$IV_c = \{e_1, e_2, \dots, e_{n_c}\} \quad (1)$$

The transmission model of cascades in this work follows the independent cascade model of [1]. In this model, each node infects each of its neighbors independently by a random variable. Propagating these information cascades over the underlying network builds the diffusion network which is called $G^* = (V^*, E^*)$. The covering percentage of diffusion network over the underlying network depends on a metric, called the diffusion rate δ . The parameter that controls how far a cascade can spread is denoted by β [1].

We call an “element set”, T , which refers to a set of diffusion network elements that could be nodes, links or cascades. For element $e \in T$, L is defined as a finite set of labels which shows a specific feature of e . We assume that the label $l_e \in L$ is assigned to each element e by a function $f : T \rightarrow L$ which is called the measurement function. For example, infection is a label for each node that shows whether

this node is infected during the diffusion process or not. The measurement function f for this label will match nodes $u \in V$ to the set $L = \{0, 1\}$ ($f(u) = 0$, if node u is not infected and $f(u) = 1$, otherwise). To measure a characteristic of network G , We consider the average function $A_G(f)$ that is defined as:

$$A_G(f) = \frac{\sum_{u \in V} f(u)}{|V|} \quad (2)$$

In the infection example, this average shows the percentage of infected nodes by the diffusion process to all the nodes of the underlying network.

B. Problem Definition

Consider the graph G as the underlying graph for a sampling approach that will yield the sampled graph $G(S)$ as a sub-graph of G . Let define the accuracy of a sampling approach as:

$$\lambda = 1 - \frac{|A_{G^*}(f) - A_{G(S)}(f)|}{A_{G^*}(f)} \quad (3)$$

Our goal is to evaluate the proposed sampling approaches (SBS and DBS), in terms of the accuracy in measuring the characteristics of diffusion process. The sampling rate μ and diffusion rate δ are the constraints of this problem.

IV. EXPERIMENTAL EVALUATION

In this section, the performance of SBS and DBS will be analyzed by measuring a number of newly defined diffusion process characteristics. In both approaches, we use two sampling techniques of BFS and RW. As SBS and DBS will be done over the underlying network and diffusion network respectively, we should consider different sampling ratios for each of these approaches to result in the same number of edges in the sampled network. Since diffusion network is a subset of the underlying network (by proportion of δ), different sampling rates (μ) from 0 to 1 for DBS over diffusion network will be equal to $0 < \mu < \delta$ for SBS over the underlying network. For easier readability, we consider the sampling rate related to DBS in all figures. For a reasonable rate of information diffusion, we also consider $\delta = 0.5$ which means G^* will cover half of the G .

A. Synthetic Dataset

In order to construct synthetic networks, two well-known models for generating such networks are used: Kronecker [18] and Forest Fire model [19]. Using different sets of parameters in the Kronecker model, we generate three different networks named Random [20], Hierarchical [21], and Core-Periphery network [22]. The parameters for generating networks and propagating cascades are provided in Table I.

B. Real Dataset

We use two real-world networks. The first network is a political blogosphere which has 1490 blogs and 19090 directed links between them [23]. The other is a co-authorship network of theory scientists that contains 2742 directed links between 1589 scientists [24].

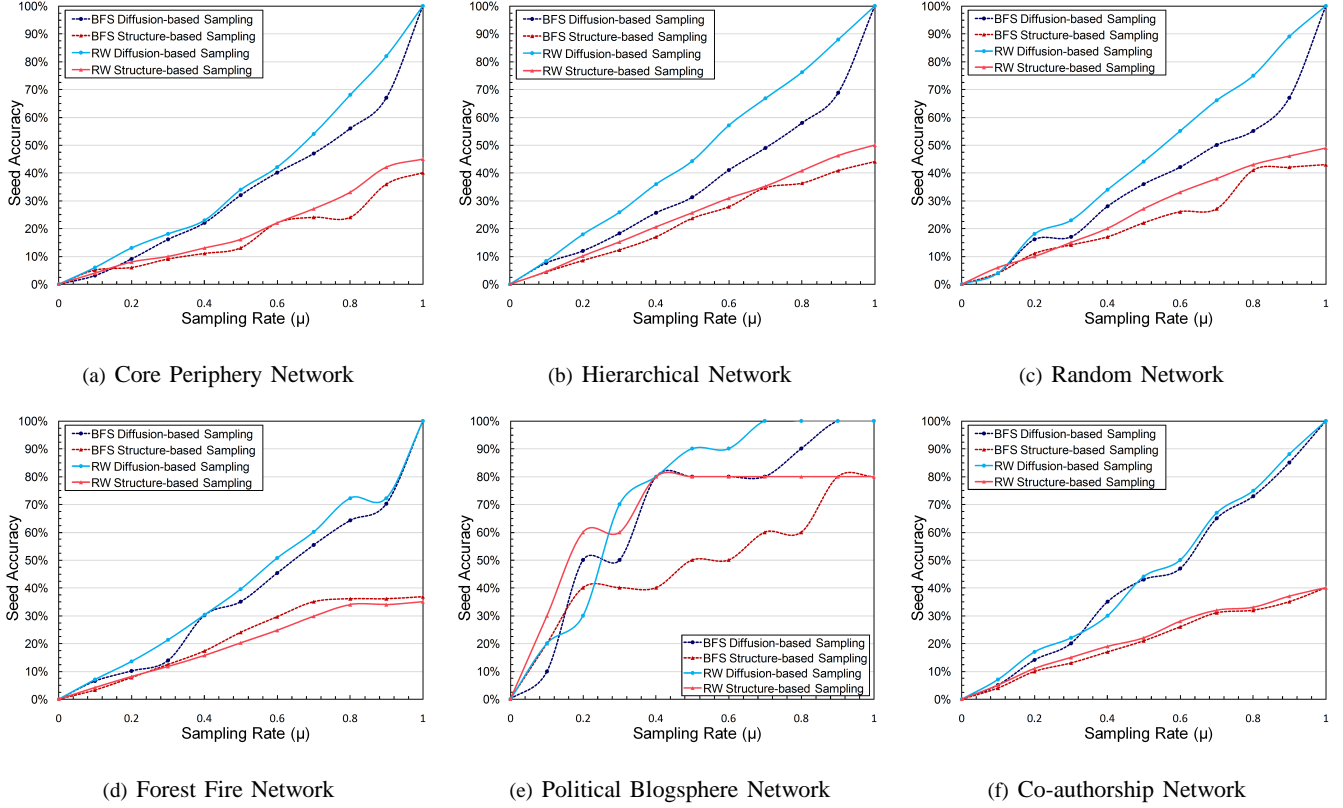


Fig. 2. Accuracy of seed characteristic measurement in different sampling approaches.

TABLE I
THE SETTING PARAMETERS.

Network	Parameter Matrix	Nodes	Edges	β
Core-Periphery	[0.9, 0.5; 0.5, 0.3]	8192	15000	0.1
Hierarchical	[0.5, 0.5; 0.5, 0.5]	8192	11707	0.5
Random(ER)	[0.9, 0.1; 0.1, 0.9]	8192	15000	0.5
Forest Fire	[5; 0.12; 0.1; 1; 0]	10000	14305	0.5

C. Diffusion Characteristics Evaluation

In previous studies [8], [9], the evaluation of diffusion process is done by measuring some characteristics which are more dependent on the structure of the network rather than its propagation behavior. In the following, we propose a number of diffusion-based characteristics and classify them into three categories. This classification can be effective in evaluation and analysis of diffusion characteristics. Additionally, the proposed characteristics cover a wide range of measures in gauging diffusion network sampling approaches.

Node-based Characteristics. The beginners of an infection process play a critical role in the diffusion process. In many applications such as political issues, starting a diffusion process is more important than continuing it. Therefore, the beginner of an infection, called “Seed”, can be considered as a node-based characteristic. We define the measurement function $f(u)$ for seed characteristic as follows: $f(u) = 1$, if node u is a seed

in the original network, and $f(u) = 0$, otherwise. The previous definitions (e.g. [9]), only consider the number of seeds in the sampled network while this new characteristic determines the common seeds between the original and sampled network as a more realistic definition.

As it was explained in section II, tracking diffusion paths in DBS approach should result in higher accuracy of analyzing diffusion process in comparison with SBS. Measuring the accuracy of seed characteristic in both SBS and DBS approaches confirms this claim. In Figure 2, the accuracy of seed characteristic measurement has been depicted in all of the synthetic and real networks. Our results show that the seed accuracy in DBS grows faster than SBS by increasing the sampling rate. In higher sampling rates, this phenomenon will result in considerable performance difference between these approaches (up to 65%).

Nevertheless, SBS in the blogosphere network can decrease this difference by up to 25% in contrast to the other networks (Figure 2(e)). This different behavior is the result of the network density. Considering the relation between the number of nodes and edges given by $E(t) \propto N(t)^a$ [19], the densification exponent (a) in the blogosphere network is more than the others (Table II). This higher density gives SBS more options in visiting the nodes to find the beginners of the infection. Therefore, SBS can achieve higher accuracy in a dense network such as blogosphere network.

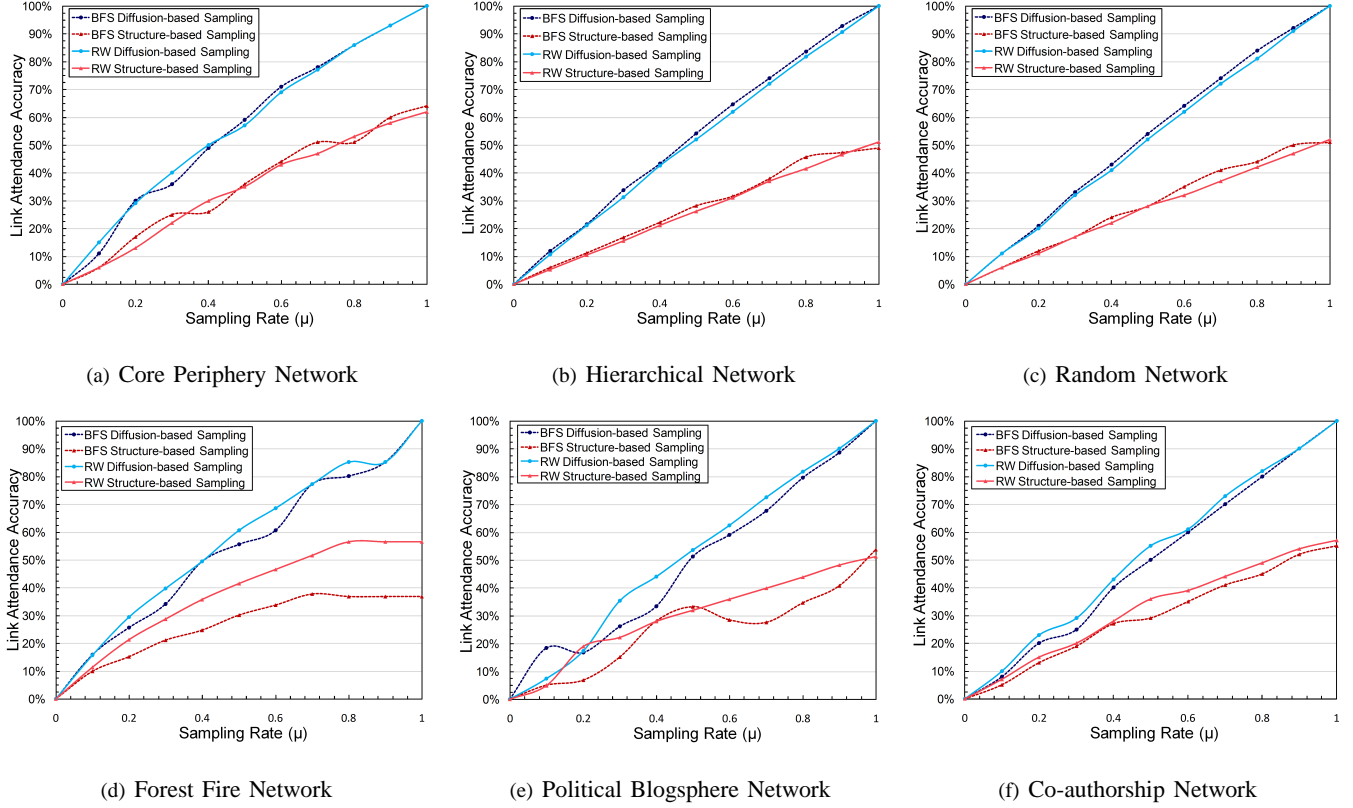


Fig. 3. Accuracy of Link Attendance characteristic measurement in different sampling approaches.

TABLE II
THE NETWORKS DENSIFICATION EXPONENT.

Network	Dens. Exp. (a)
Core-Periphery	1.06
Hierarchical	1.03
Random(ER)	1.06
Forest Fire	1.03
Blogshpere	1.34
Co-Authorship	1.07

Link-based Characteristics: In the diffusion process, some links have more attendance than the others. These links are significant in some applications such as finding potential paths of infection propagation in the epidemic spreading [12]. Let $C_e = \{c | e \in IV_c\}$ be the set of cascades in which link e appears. We define the “Link Attendance” characteristic by the measurement function $f(e)$ for link e as $f(e) = |C_e|$. As shown in Figure 3, we can obtain more link attendance accuracy with DBS compared to SBS. This performance gap will be greater in higher sampling rates in a manner similar to the seed characteristic.

Cascade-based Characteristics: In general, the depth of an infection can be determined by the diffusion path length. Since the diffusion network is usually assumed to be a tree, the depth characteristic is defined by the length of the tree [8], [9]. However, a real diffusion network is not a tree. Therefore,

we consider the length of a cascade, c , to define the depth characteristic by the measurement function $f(c) = |IV_c|$.

As Figure 4 shows, SBS can achieve higher accuracy in depth characteristic compared to the seed and link attendance characteristics. This is the result of inherent difference between these characteristics. More specifically, seed and link attendance are individual-based characteristics while depth is related to the cascades as a group-based characteristic. This feature of the depth characteristic gives SBS approach more choices in exploring the underlying network. Therefore, the performance difference between SBS and DBS will be decreased for depth accuracy from 60% to 35%, in average.

D. Discussion

Here, we address the general superiority of DBS vs. SBS by considering the effect of different sampling rates. The sampling rate has been divided to three ranges; (1) low range: $0 < \mu \leq 0.3$, (2) medium range: $0.3 < \mu \leq 0.6$, and (3) high range: $0.6 < \mu \leq 1$. Considering these sampling ranges, we first measure the average accuracy of each characteristic for each sampling approach, in all networks. Then we illustrate the superiority of DBS over SBS by calculating their performance difference (refer to Figure 5). Although DBS performs much better than SBS in the medium and high sampling ranges (in average by 16% and 29%, respectively), the performance difference between them is about 7% in the low sampling rates.

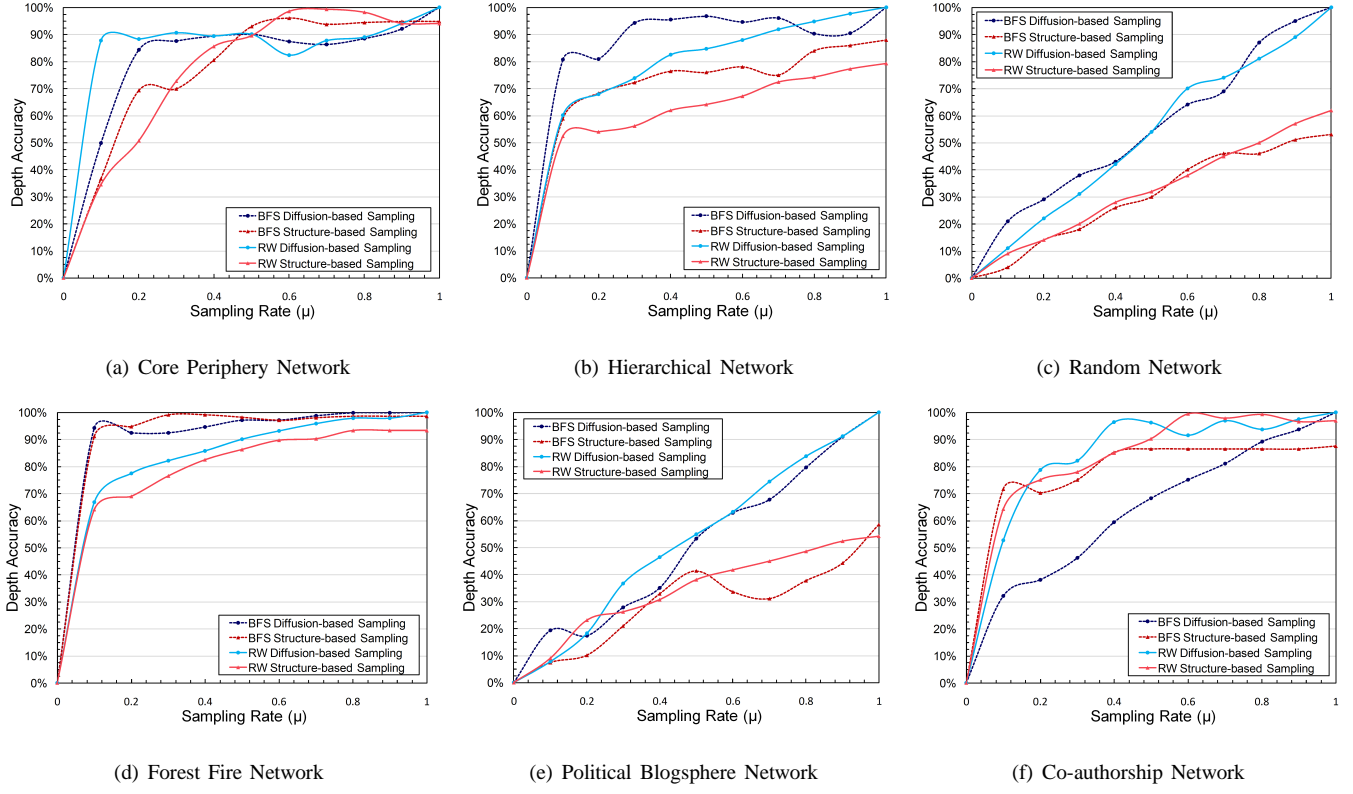


Fig. 4. Accuracy of depth characteristic measurement in different sampling approaches.

Therefore, in real large scale systems that we have to sample the network with a low sampling rate, SBS would be a better choice because of its lower time complexity in collecting data, compared to DBS.

Moreover, we investigated the performance of two sampling methods of RW and BFS for both SBS and DBS approaches. Figure 6 illustrates the superiority of RW with respect to BFS in measuring diffusion characteristics for different sampling rates. However, the performance difference of RW and BFS techniques in average is about 3% which is much lower than the performance difference between SBS and DBS approaches.

This fact shows the sampling approaches (namely DBS and SBS) are more important than the techniques which are used to implement them (namely RW and BFS).

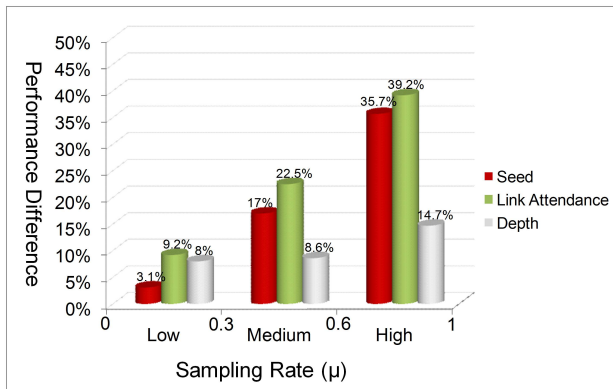


Fig. 5. Diffusion-based Sampling (DBS) vs. Structure-based Sampling (SBS)

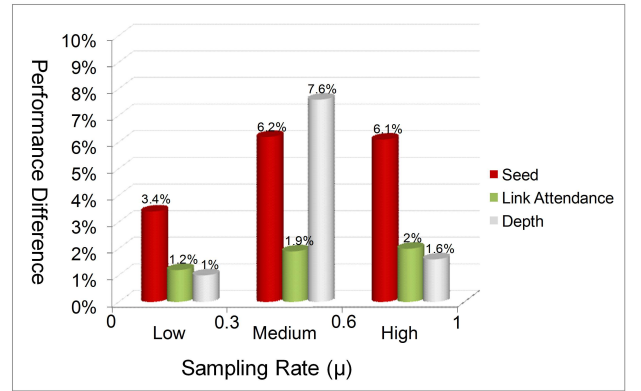


Fig. 6. Random Walk (RW) vs. Breadth-First Search (BFS)

V. CONCLUSIONS

In this paper, we introduced the “Structure-based Sampling (SBS)”, and “Diffusion-based Sampling (DBS)” approaches for the analysis of information diffusion networks. These approaches were evaluated over large synthetic and real networks in terms of the newly proposed diffusion characteristics. Our experiments showed that tracking diffusion paths with DBS

approach will result in more accurate analysis of diffusion process in comparison with SBS. In addition, by increasing the sampling rates more accurate results are achieved in measuring seed and link attendance characteristics by using DBS. However, a cascade-based characteristic has different behavior compared to the node-based and link-based characteristics. In this case, the performance difference between SBS and DBS in measuring cascade-based characteristics accuracy is decreased. Furthermore, our analysis on the performance of the introduced sampling approaches showed that structure-based sampling is preferable in the large scale systems in which low sampling rates are more feasible. Moreover, we have found that the sampling techniques such as RW and BFS are less significant than the sampling approaches (i.e. DBS and SBS) on analysis of the diffusion process.

We believe that our results provide a promising step towards understanding the sampling approaches in analysis and evaluation of diffusion processes. There are several interesting directions for future work. Proposing a new sampling approach which can decrease the gap between structure-based and diffusion-based sampling approaches is one of our main future goals. Including other diffusion aspects such as infection times to define new diffusion characteristics is another aim which we would consider in the future.

VI. ACKNOWLEDGMENTS

This research has been partially supported by ITRC (Iran Telecommunication Research Center) under grant number 6479/500 (90/4/22).

REFERENCES

- [1] M. Gomez-Rodriguez, J. Leskovec and A. Krause, *Inferring networks of diffusion and influence*, In proc. of KDD '10, pages 1019-1028, 2010.
- [2] Twitter Blog: \neq numbers. Blog.twitter.com., Retrieved 2012-01-20, <http://blog.twitter.com/2011/03/numbers.html>.
- [3] M. Gjoka, M. Kurant, C. T. Butts and A. Markopoulou, *Practical Recommendations on Crawling Online Social Networks*, IEEE J. Sel. Areas Commun, 2011.
- [4] M. Salehi, H. R. Rabiee, N. Nabavi and Sh. Pooya, *Characterizing Twitter with Respondent-Driven Sampling*, International Workshop on Cloud and Social Networking (CSN2011) in conjunction with SCA2011, No. 9, Vol. 29, pages 5521-5529, 2011.
- [5] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel and B. Bhattacharjee, *Measurement and analysis of online social networks*, Proceedings of the ACM SIGCOMM conference on Internet measurement, pages 29-42, 2007.
- [6] J. Leskovec and Ch. Faloutsos, *Sampling from large graphs*, Proceedings of the ACM SIGKDD conference on Knowledge discovery and data mining, pages 631-636, 2006.
- [7] M. Salehi, H. R. Rabiee, and A. Rajabi, *Sampling from Complex Networks with high Community Structures*, Chaos: An Interdisciplinary Journal of Nonlinear Science , 2012.
- [8] D. Liben-Nowell and J. Kleinberg, *Tracing information flow on a global scale using Internet chain-letter data*, Proc. of the National Academy of Sciences, 105(12):4633-4638, 25 Mar, 2008.
- [9] M. D. Choudhury, Y. Lin, H. Sundaram, K. S. Candan, L. Xie and A. Kelliher, *How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?*, Proc. of ICWSM , 2010.
- [10] E. Sadikov, M. Medina, J. Leskovec and H. Garcia-Molina, *Correcting for missing data in information cascades*, WSDM, pages 55-64, 2011.
- [11] M. Gomez-Rodriguez, D. Balduzzi and B. Scholkopf, *Uncovering the Temporal Dynamics of Diffusion Networks*, Proc. of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.
- [12] M. Eslami, H. R. Rabiee and M. Salehi, *DNE: A Method for Extracting Cascaded Diffusion Networks from Social Networks*, IEEE Social Computing Proceedings, 2011.
- [13] M. Eslami, H. R. Rabiee and M. Salehi, *Diffusion-Aware Sampling and Estimation in Information Diffusion Networks*, IEEE Social Computing Proceedings, 2012.
- [14] C.X. Lin, Q. Mei, Y. Jiang, J. Han and S. Qi, *Inferring the Diffusion and Evolution of Topics in Social Communities*, SNA KDD, 2011.
- [15] Ch. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy and B. Y. Zhao, *User interactions in social networks and their implications*, EuroSys '09: Proceedings of the 4th ACM European conference on Computer systems, pages 205-218, 2009.
- [16] L. Lovas, *Random walks on graphs: a survey*, Combinatorics, 1993.
- [17] M.R. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork, *On near-uniform URL sampling*, Proceedings of the World Wide Web conference on Computer networks, pages 295-308, 2000.
- [18] J. Leskovec and C. Faloutsos, *Scalable modeling of real graphs using Kronecker multiplication*, Proc. of ICML, pages 497-504, 2007.
- [19] J. Leskovec, J. Kleinberg and C. Faloutsos, *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations*, Proc. of KDD , 2005.
- [20] P. Erdős and A. Rnyi, *On the evolution of random graphs*, Publ. Math. Inst. Hung. Acad. Sci., 5: page 17, 1960.
- [21] A. Clauset, C. Moore and M. E. J. Newman, *Hierarchical structure and the prediction of missing links in networks*, Nature, 453: pages 98-101, 2008.
- [22] J. Leskovec, K.J. Lang, A. Dasgupta and M.W. Mahoney, *Statistical properties of community structure in large social and information networks*, WWW, pages 695-704, 2008.
- [23] L.A. Adamic and N. Glance, *The political blogosphere and the 2004 US Election*, Proc. of the WWW-2005 Workshop on the Weblogging Ecosystem, 2005.
- [24] M. E. J. Newman, *Finding community structure in networks using the eigenvectors of matrices*, Preprint physics/0605087, 2006.